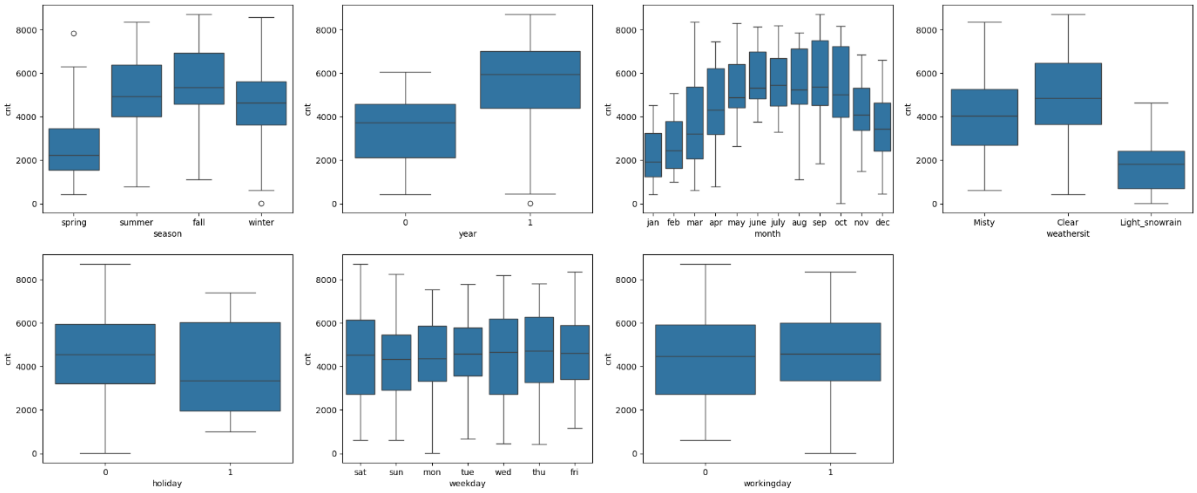
# Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

**Answer:**

Please find the inference from my analysis on the Categorical variables namely- season, year, month, weathersit, holiday, weekday and working day. The inferences are based on the below plots from the python notebook:



Inference:
There seems to be high demand for bike in fall season

- The bike demand seems to be increasing per year with May to October month having particularly high demand
- Bike rental demand increases when the weather is clear
- Median Bike rentals decrease when there is a holiday, implying customers are using rental bike for office commute
- Median Bike rentals seem to be same for working and non-working days

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

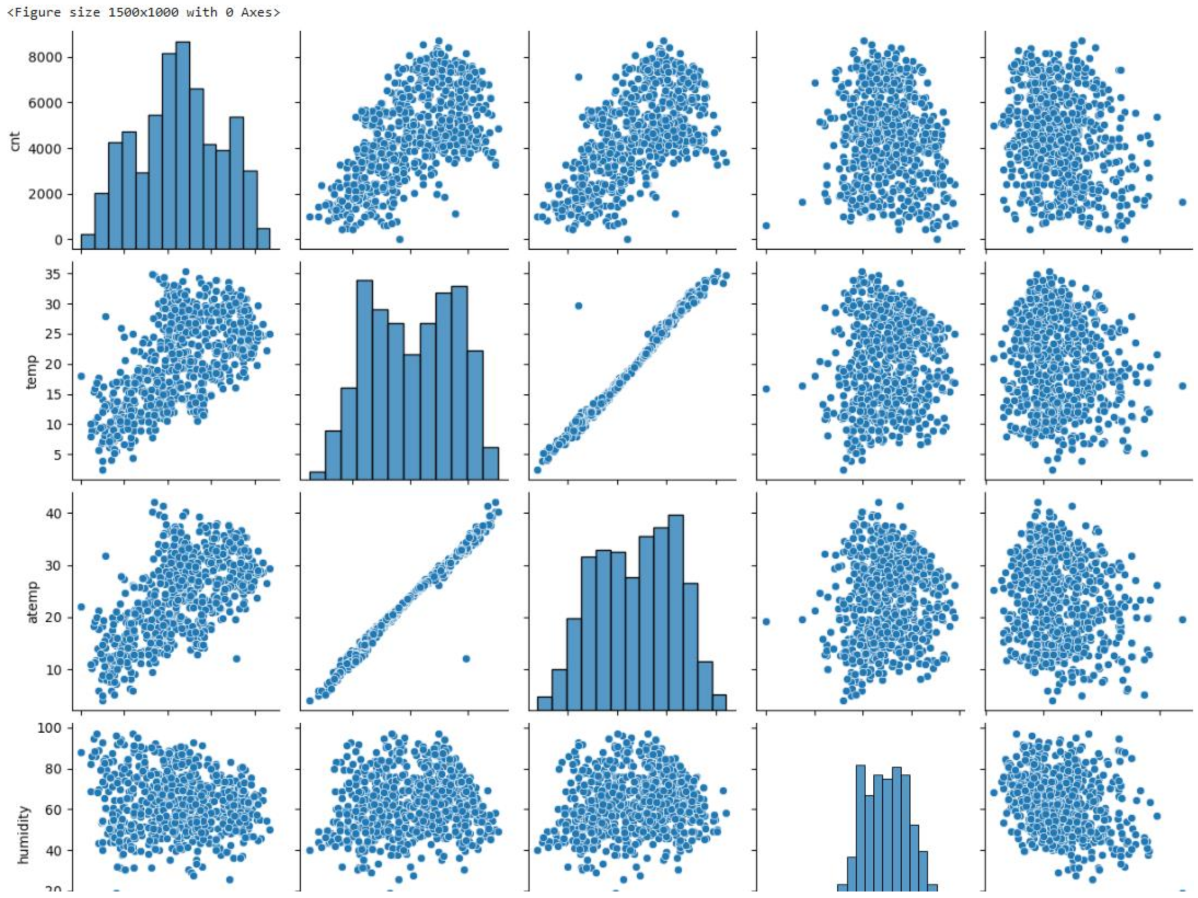
**Answer:**

Dummy variables are created for those categorical variables which has more than 2 levels. If we create dummy variables for all the levels of the category, then there would be high multicollinearity between the dummy variables, since one dummy variable can be explained by the absence of all other dummy variables. Hence, we need to use drop_first=True. For Ex:- We have a categorical variable- season which has 4 values- spring, summer, fall and winter. If we have 3 dummy variables namely spring, summer and winter, the absence of these 3 (0 value in these 3) would mean that season is fall. We don't need to have fall as a dummy variable. If we include, then we will get high VIF values (INF) due to multicollinearity.

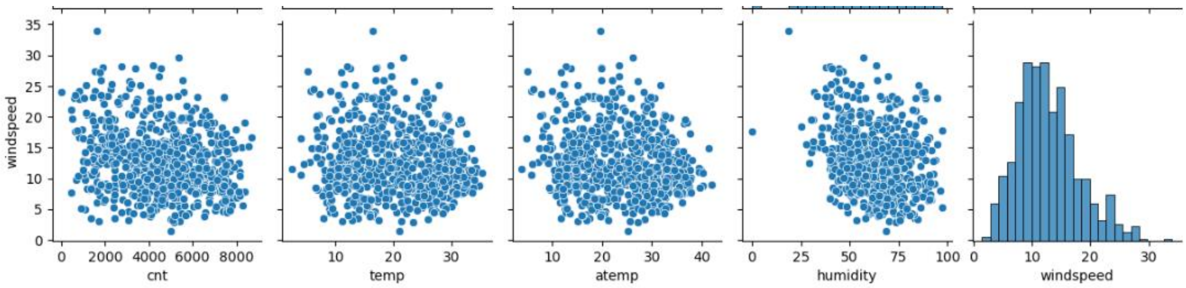
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

**Answer:**

Both 'temp' and 'atemp' are having the highest correlation with the target variable- cnt. PFB the pair-plot for reference:



```
<Figure size 1500x1000 with 0 Axes>
```

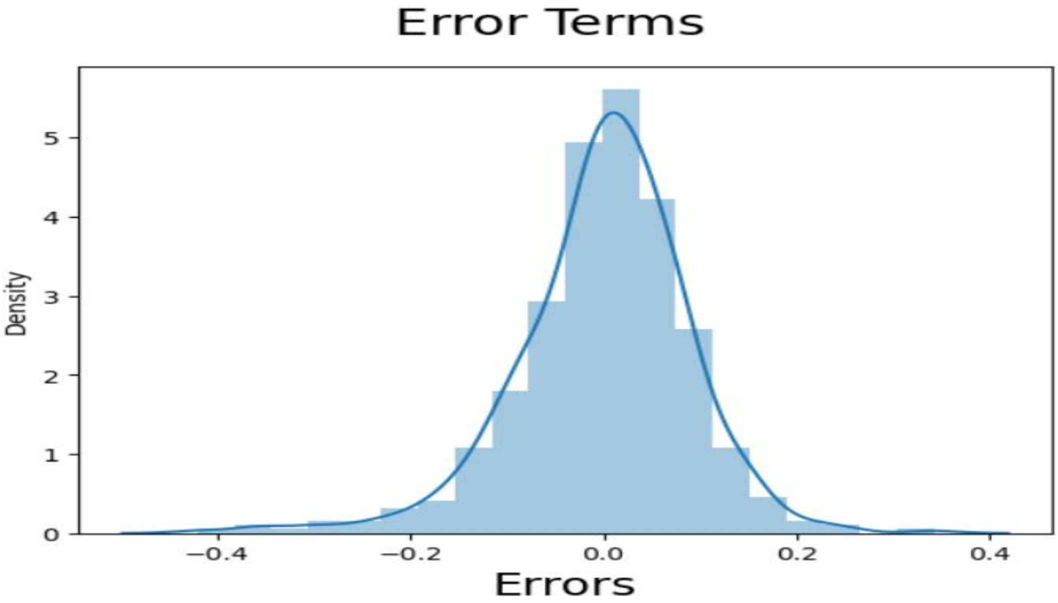


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

**Answer:**

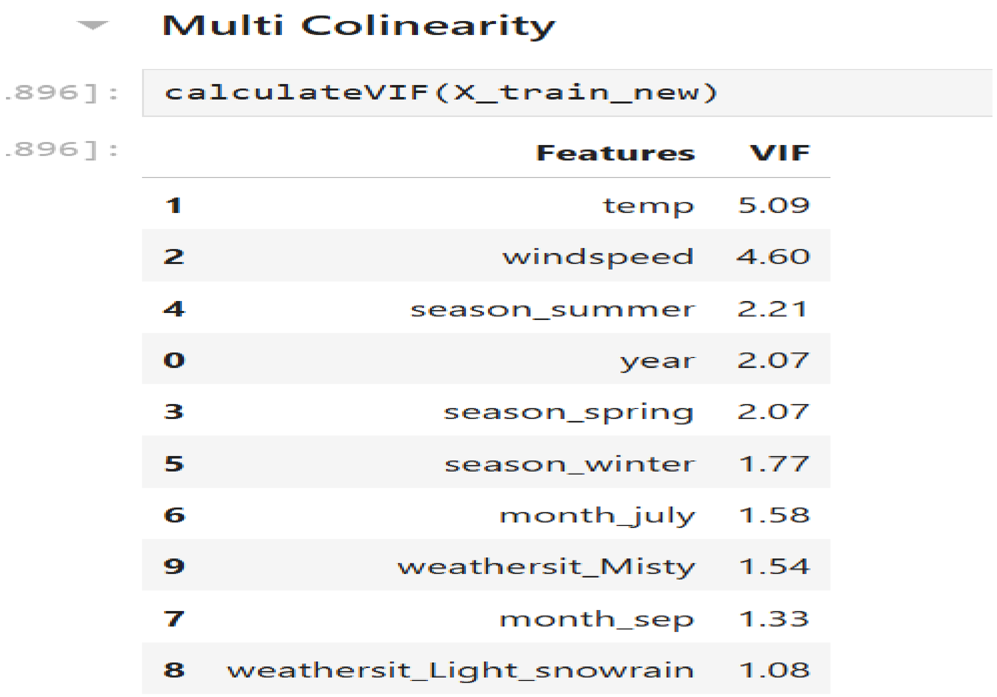
I have validated the assumption of Linear Regression Model based on the below checks –

- Normality of error terms
  - Error terms should be normally distributed



  - As we can see from above, error terms are normally distributed

- Multicollinearity check
  - There should be insignificant multicollinearity among variables.



## Multi Colinearity

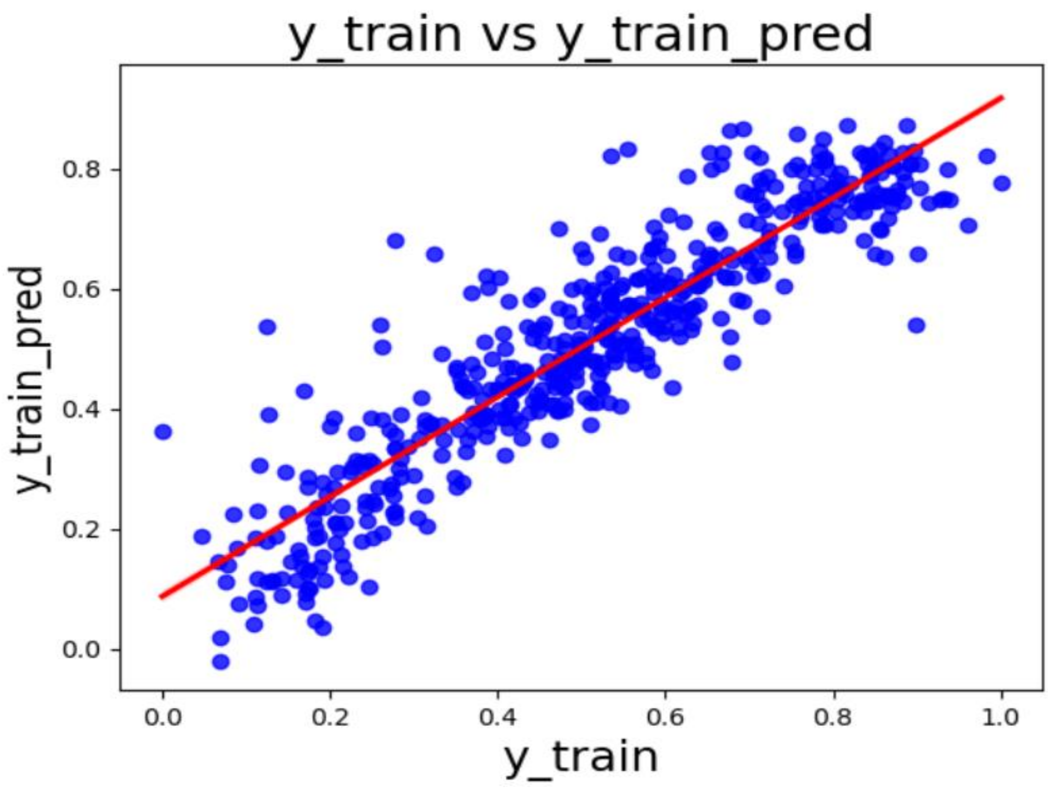
```
.896]:  calculateVIF(X_train_new)
```

.896]:

|   | Features | VIF |
|---|---|---|
| 1 | temp | 5.09 |
| 2 | windspeed | 4.60 |
| 4 | season_summer | 2.21 |
| 0 | year | 2.07 |
| 3 | season_spring | 2.07 |
| 5 | season_winter | 1.77 |
| 6 | month_july | 1.58 |
| 9 | weathersit_Misty | 1.54 |
| 7 | month_sep | 1.33 |
| 8 | weathersit_Light_snowrain | 1.08 |

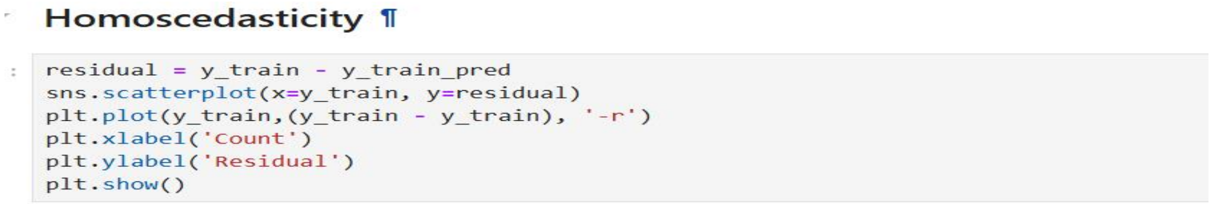
- VIF values are less than 5 which shows there is low multicollinearity between features

- Linear relationship validation
  - Linearity should be visible among variables



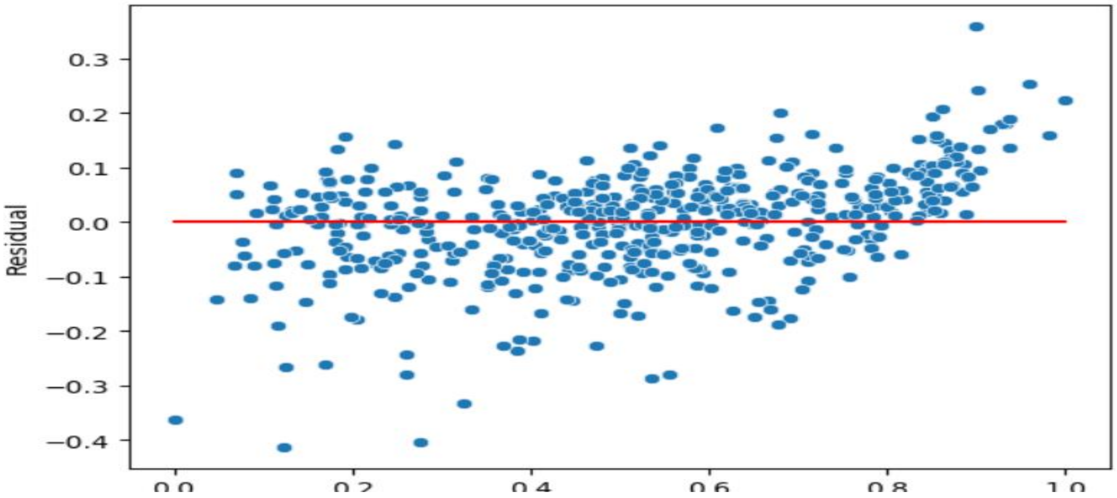
- As seen from above plot, there is a clear linear relationship

- Homoscedasticity
  - There should be no visible pattern in residual values.



Homoscedasticity ¶

```
residual = y_train - y_train_pred
sns.scatterplot(x=y_train, y=residual)
plt.plot(y_train,(y_train - y_train), '-r')
plt.xlabel('Count')
plt.ylabel('Residual')
plt.show()
```



  - As seen from above plot, the shape is consistent and there is no funnel shape showing that error terms are spread constantly across all values of target variable

- No Auto-correlation
  - Durbin-Watson value of final model lm6 is 1.994, which signifies there is no autocorrelation.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

   **Answer:**

   The Top 3 features contributing significantly towards the demand of shared bikes are:

   Temperature, Year and Weather- Light_snowrain

## General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

**Answer:**

Linear Regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with a set of independent variables, linear meaning that a change in independent variable will lead to a change in the dependent variable and this change will be a linear function (unlike others like exponential etc.).

The Mathematical relationship for Simple Linear regression can be expressed as:

> Y=mX+C, where Y is the dependent or target variable, X is the independent variable, m is the slop of the regression line and C is a constant also knows as the intercept

The linear relationship can be positive or negative. There are 2 types of linear regression, based on the number of independent variables:

- **Simple Linear Regression**: Here, there is 1 dependent and 1 independent variable. It can be represented as below:

  $$\mathrm{y}= b_o + b_1 x$$

  Here, bo is the intercept, b1 is coefficient or slope, x is the independent variable and y is the dependent variable.

- **Multiple Linear Regression**: There can be many independent variables. It can be represented as below:

  $$\mathrm{y}= b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots . + b_n x_n$$

  Here, bo is the intercept, b1,b2,b3,b4…,bn are coefficients or slopes of the independent variables x1,x2,x3,x4…,xn and y is the dependent variable.

The main aim of a linear regression model is to fit a line across the data points that can explain or predict the value of the target variable.

There are few algorithms that are used for above in order to minimise the difference between the predicted and actual values of the target variable.

- **OLS (Ordinary least Squares) algorithm**: In this method, we calculate the sum of the Squares of the error on all data points, also called RSS (Residual Sum of Squares) and the line which has the least value for this sum is used as the best fit line or model

  $$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_i)^2$$

- **R2 or Co-efficient of Determination**: R2 is a score which explain what portion of the variance is explained by the model. The higher this number, the better the model is. The Equation of R2 is as follows:

$$R2 = 1 - RSS/TSS$$

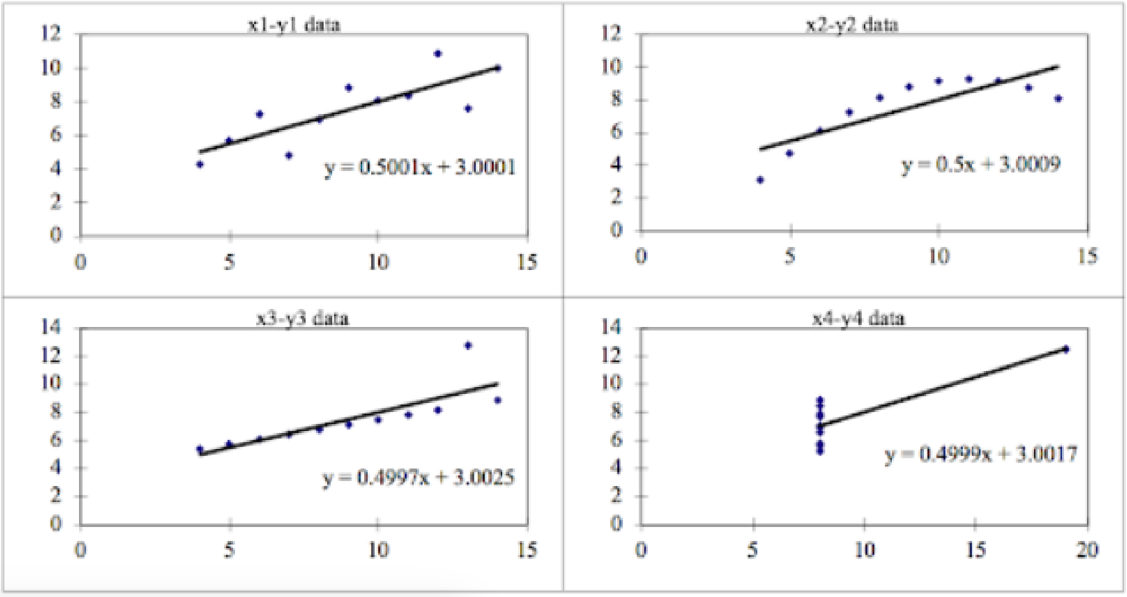
Where RSS is the Residual Sum of Squares and TSS is Total Sum of Squares

$$TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

2. **Explain the Anscombe's quartet in detail. (3 marks)**

## Answer:

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics in terms of mean, variance, R-squared, Correlations, and linear regression lines, but having different graphical representations on a scatter plot.



Anscombe's Quartet Four Datasets

- **Data Set 1:** fits the linear regression model pretty well.
- **Data Set 2:** cannot fit the linear regression model because the data is non-linear.
- **Data Set 3:** shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- **Data Set 4:** shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

Anscombe's Quartet helps us to understand the importance of data visualization. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

3. **What is Pearson's R? (3 marks)**

## Answer:

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's R, the Pearson product-moment correlation coefficient (PPMCC), or Bivariate correlation. It is a statistic that measures the linear correlation between two variables.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

The value of the Pearson's R is between -1 to +1. When the correlation coefficient comes down to zero, then the data is said to be not related. While, if we are getting the value of +1, then the data are positively correlated and -1 has a negative correlation.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

## Answer:

Scaling is the process of bringing all the numerical variables of a dataset on scale with each other. This is so that the coefficients obtained as part of the linear regression modelling do not lose significance due to high numerical value of certain variables.

Mainly two types of scaling are performed:

- **Normalization**: Here, the variables are scaled such that all the values lie between 0 and 1. This is also known as MinMax scaling. Please find below formula for Normalized or MinMax scaling:

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

- **Standardization**: Here, the variables are scaled in such a way that the mean is 0 and Standard Deviation is 1. Please find below formula for Standardized scaling:

$$x = \frac{x - mean(x)}{sd(x)}$$

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

## Answer:

VIF (Variance Inflation Factor) is a way of measuring multicollinearity. Its formula is as below:

VIF = 1/(1-R2), where R2 is the R-Squared score of the model with the rest of the variables of the model as independent variables.

So, larger the R2 means that the variable can be predicted mostly using the other variables in the set, i.e. variable is having high multicollinearity.

A Value of INF means that the R2 is 1, which may be the case when there 2 sets of variables with a 1:1 relationship. In such cases, we need to identify and use only one such variable in the model.

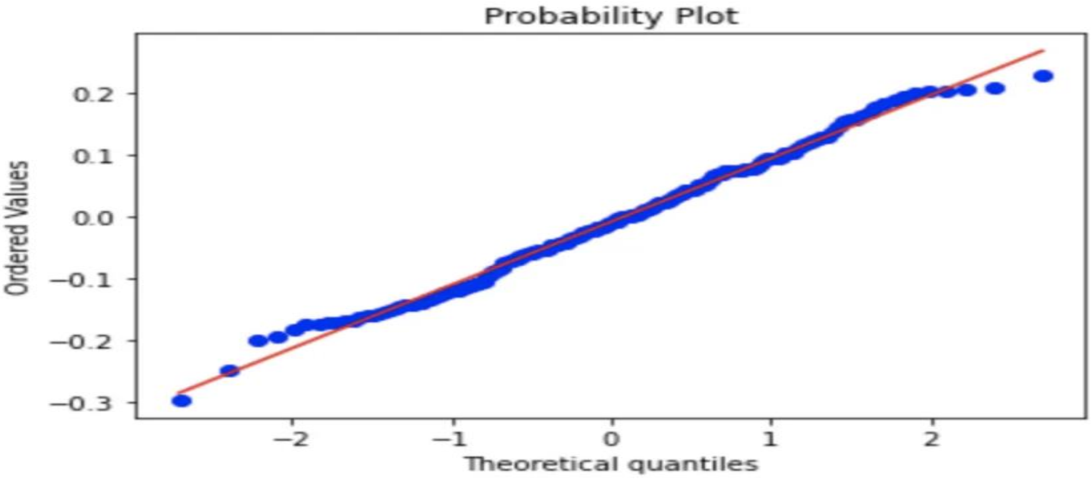
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks**

## Answer:

Q–Q plot or Quantile-Quantile plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. It is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

It is a scatter plot created by plotting a set of two quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.



**Importance of QQ Plot in Linear Regression:**

In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.

Advantages:

- It can be used with sample size also

- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot Q-Q plot use on two datasets to check
- If both datasets came from population with common distribution
- If both datasets have common location and common scale
- If both datasets have similar type of distribution shape
- If both datasets have tail behaviour